Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 38 (2012) 1072 – 1076

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Modeling, Optimization and Computing (ICMOC-2012)

Empirical Comparison of Sampling Strategies for Classification

Kaberi Das[a], Prem Pujari Pati[b], Debahuti Mishra[c], Lipismita Panigrahi[d]

[a,b,c,d]*Institute of Technical Education and Research, Siksha O Anusandhan Deemed to be University, Bhubaneswar, Odisha, India*

Abstract

Data sets contain very large amount of data which is not an easy task for the user to scan the entire data set. The researcher's initial task is to formulate a rational justification for the use of sampling in his research. Sampling has been often suggested as an effective tool to reduce the size of the dataset operated at some cost to accuracy. It is the process of selecting representatives which indicates the complete data set by examining a fraction. Due to sampling we overcome the problems like; i) in research it is not possible to collect and test each and every element from the data base individually; and ii) study of sample rather than the entire dataset is also sometimes likely to produce more reliable results. This paper focuses on different types of sampling strategies applied on neural network. Here sampling technique has been applied on two real, integers and categorical dataset such as yeast and hepatitis data set prior to classification. The main objective of this paper is an empirical comparison of different sampling strategies for classification which gives more accuracy.

Keywords: Data Reduction; Simple Random Sampling; Cluster Sampling; Neural Network

1. Introduction

Data reduction techniques [1] are approaches in charge of diminishing the amount of information in order to condense both memory as well as execution time. It is an important aspect of data preparation, which seeks to reduce large databases to convenient sizes, also assist judgment makers to know the exact dimensionality of its business database. Data reduction [2] obtains a reduced representation of the data set which is much smaller in volume, yet produces the same systematic results. There are a number of strategies for data reduction which include data aggregation, dimension reduction and data compression. These sampling is used as a data reduction technique which helps the user to extract the information quickly. Sampling [3] will enable the researchers to collect a smaller amount of data that represent the whole group which will save time, money and other resources, while not compromising on reliability of information. Towards data reduction, sampling [3-4] is most commonly used to estimate the answer to an aggregate query. It can be categorized into two: probability sampling and non-probability sampling. Probability sampling technique involves random selection while non-probability sampling does not involve random selection. The rest of the paper is organized as follows; section 2 deals with related work on the sampling strategies, in section 3 the types of sampling and neural network has been discussed, section 4 shows the proposed model, in section 5 the experimental result has been illustrated and section 6 gives the conclusion and future work.

* Corresponding author. Tel.: +91-8895485778; fax: +91-674-2351880.
E-mail address: prempati016@gmail.com

1877-7058 © 2012 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2012.06.135

2. Related Work

Punam V. Khandar et al. [3] presented on knowledge discovery techniques by improving approach with efficiently identifying the sampling dataset and data mining algorithm to mine the sampled data. V. Umarani et al. [4] presented an overview of existing sampling based association rule mining algorithms and shows how different strategies are in a specific data mining task given specific data sets in order to provide users a set of guidelines for them to make decisions on which context it will be more suitable to use which sampling strategy. Basel et al. [5] recently presented a parameterized sampling algorithm for association rule mining which extracts sample datasets based on certain parameters which gives 98% accuracy. Venkatesan et al. [6] proposed a different view of analyzing the quality of solution by theoretical framework which gives a comprehensive explanation for well known empirical success of sampling for association rule mining. Doug Stewart [7] recently suggested that random sampling of e-Discovery data sets yields results consistent with well established statistical principles. It shows that Simple Random Sampling (SRS) can be used to accurately make predictions about the composition of e-Discovery data sets and thus validate e-Discovery processes.

3. Preliminaries

Sampling [3][6] is one of the important and popular data reduction techniques that are used to mine huge volume of data efficiently. The different types of sampling strategies are explained below.

3.1 Probability Sampling

Probability samples are the only type of samples where the results can be generalized from the sample to the whole [4]. It tends to be more difficult and costly to conduct. In addition, it allows the researcher to calculate the precision of the estimates obtained from the sample and to specify the sampling error. Four basic types of methodologies are most commonly used for conducting probability samples; these are simple random, stratified, cluster and systematic sampling out of which only simple random sampling and cluster sampling is taken.

(a) *Simple Random Sampling*: Here the sample is drawn in such a way that each data has an equal chance of being drawn during each selection round. Samples may be drawn with or without replacement [4].In practice, however, most simple random sampling for survey research is done without replacement; i.e. an item selected for sampling is removed from the dataset for all subsequent selections.

(b) *Cluster Sampling*: Cluster sampling [5] is similar to stratified sampling because the dataset to be sampled is subdivided into mutually exclusive groups. In cluster sampling the groups are defined so as to maintain the heterogeneity of the dataset. It is the researcher's goal to establish clusters that are representative of the dataset as a whole, although in practice this may be difficult to achieve. After the clusters are established, a simple random sample of the clusters is drawn and the members of the chosen clusters are sampled.

3.2 Neural Network (NN)

It is a set of connected input/output units [8] where each connection has a weight associated with it. It learns by adjusting the weights so as to be able to correctly classify the training data. Total input to hidden layer $input = \sum_{j=1}^{n} I_j W_j$ where I_j is: the individual input values and w_j is the individual weights. Output of hidden layer $output = \frac{1}{1+e^{-input}}$, where, $\frac{1}{1+e^{-input}}$ is an activation function. One of the most commonly used techniques for learning in neural networks is called back propagation. In order for the weights of the NN connections to be adjusted, an error first needs to be calculated between the predicted response and the actual response. The number of outputs of the NN is equal to the number of inputs. Fig.1 shows the schematic representation of NN.

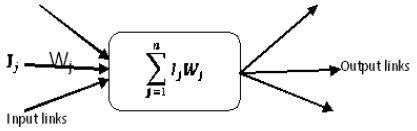


Fig.1. Schematic representation of NN

4. Schematic Representation of Proposed Model

Proposed model consists of data reduction techniques for gene expression data set which uses sampling. It is used to reduce the size of data available and normalize the data set. 70 % of normalized data is sent to the neural network classifier for training and 30% is kept for testing. On the normalized data set simple random sampling and cluster sampling have been applied. Result of sampling is then applied on NN classifier. Finally, the accuracy can be checked with the help of test data and compare the result by indicating that which sampling technique gives the best result for NN.

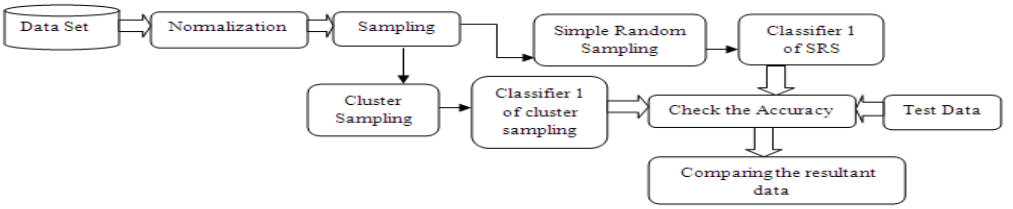
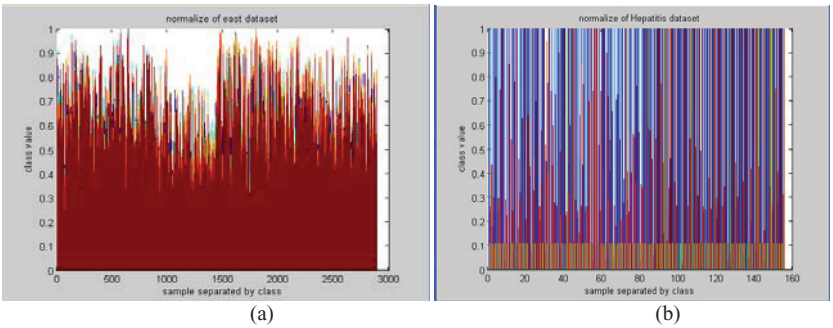


Fig.2. Schematic representation of proposed model

5. Experimental Evaluation and Result Analysis

Step1: Collection of Data Sets: The yeast [9] and hepatitis data set [9-10] have been used for experimental evaluation. The hepatitis data set contains 155 samples belonging to two different target classes out of which 109 have been used for training purpose and 46 for testing. There are 19 features, 13 binary and 6 attributes with 6–8 discrete values. Similarly yeast data set contains 1484 no. of instances from which 1039 is used for training and 445 are kept for testing purpose.

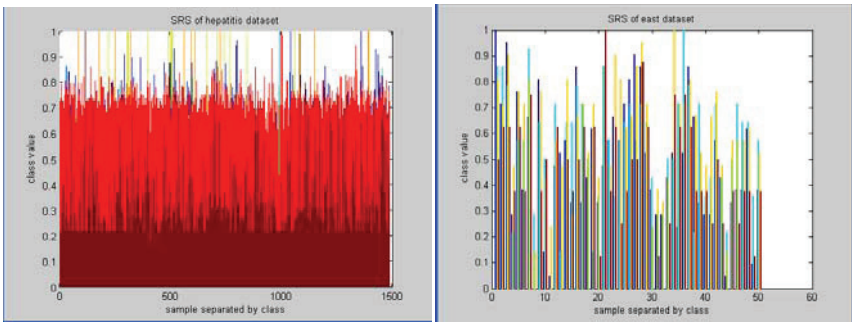
Step2: Normalization of Data Set: Dataset has multivariate characteristics and the attributes have real characteristics. Applying the min-max normalization technique the data set has been normalized which is given below. Fig. 3(a) shows the normalization of yeast data set and fig. 3(b) shows the normalization of hepatitis data set.



(a) (b)

Fig. 3. : (a) Normalization of yeast data set; (b) Normalization of hepatitis data set

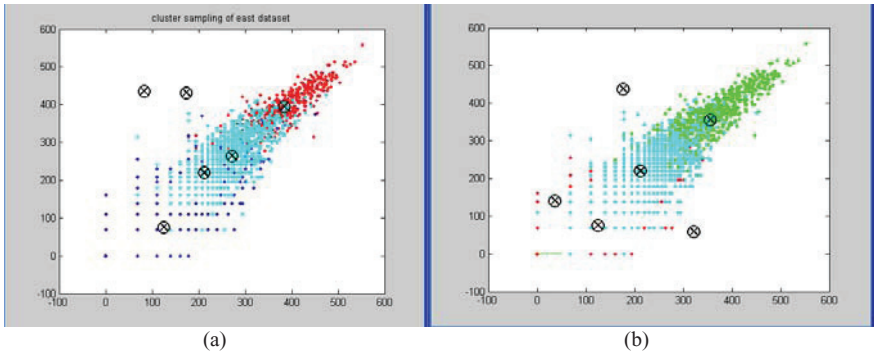
Step3: Simple Random Sampling: After the data has been normalized simple random sampling and cluster sampling are applied on the normalized data to get sample data. In simple random sampling the data are chosen randomly from both the data set and graph is plotted by taking the sample data. Fig. 4(a) and fig. 4(b) shows the result of simple random sampling on the yeast and hepatitis data set.



(a) (b)

Fig. 4. (a) Simple random sampling of hepatitis data set; (b) Simple random sampling of yeast data set

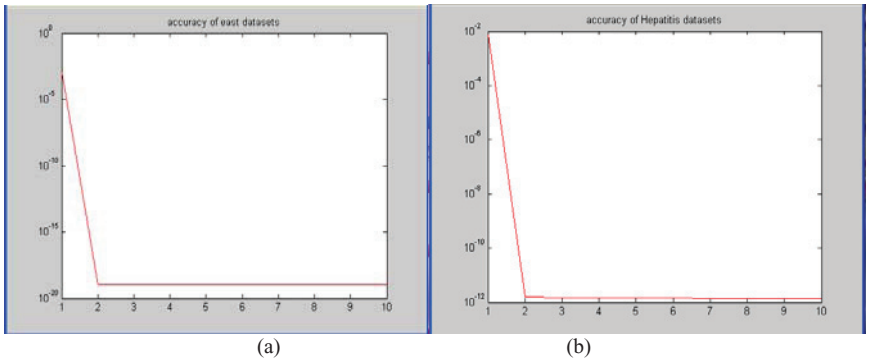
Step4: Cluster Sampling: In cluster sampling the cluster is prepared by applying clustering algorithm. After getting the clustered data again simple random sampling has been applied on those clusters and the graph has been plotted. Fig. 5(a) and fig. 5(b) shows the result of cluster sampling of hepatitis and yeast data set.



(a) (b)

Fig.5. (a) Cluster sampling of hepatitis data set; (b) Cluster sampling of yeast data set

Step5: Accuracy Measure: After finding the resultant sample data from simple random sampling and cluster sampling finally the accuracy has been checked with the help of NN classifier as shown in fig. 6(a) for yeast data set and fig. 6(b) for hepatitis data set. The graph indicates that the cluster sampling of hepatitis data set has been more accurate rather than the yeast data set.



(a) (b)

Fig.6: (a) accuracy of yeast data set; (b) accuracy of hepatitis data set

Table 1 Comparison of classification result for Yeast and Hepatitis data set with NN

Data set	**Accuracy by using NN (%)**
Yeast Data set of SRS	68.8
Yeast Data set of CS	90.1
Hepatitis Data set of SRS	77.9
Hepatitis Data set of CS	94.8

6. Conclusion and Future Work

From the experimental result we concluded that cluster sampling gives more accuracy for both the data sets in case of neural network. Data reduction is concerned with reducing the amount of data while retaining its vital characteristics. Although sampling provides a common approach which scales well and offers more flexibility, it is an essential job to know how different sampling strategies can be applied for specific data set in order to provide users to follow set of rules for making decisions. The goal of this paper indicates various sampling strategies, applied on two different datasets and check the accuracy through the classifier. For more preciseness, future work can include with extending the other sampling techniques, its selection criteria and find the efficiency using different classifiers.

References

[1] Nhien An Le Khac, Martin Bue, M-Tahar Kechadi. Studying the Impact of Partition on Data Reduction for Very Large Spatiotemporal Datasets. *The Third International Conference on Advances in Databases, Knowledge and Data Applications* 2011; 115-118.
[2] Erendira Rendon, J. Salvador Sanchez, Rene A. Garcia, Itzel Abundez, Citlalih Gutierrez and Eduardo Gasca. Data Reduction Method for Categorical Data Clustering. *LNAI 5290*, 2008; 143–152.
[3] Punam V. Khandar, Sugandha V. Dani. Knowledge discovery and sampling techniques with data mining for identifying trends in datasets. *International Journal on Computer Science and Engineering* 2010; 975-3397.
[4] V. Umarani, Dr. M. Punithavalli. Sampling based Association Rules Mining. *(IJCSE) International Journal on Computer Science and Engineering* 2010; 314-318.
[5] Basel. A new sampling technique for Association rule mining. *Journal of information science* 2009; 358-376
[6] Venkatesan T, Vinayaka Pandit, Yogish Sabharwal. Analysis of sampling techniques for Association rule mining. ACM, ICDT: 2009.
[7] Doug Stewart Daegis. Application of Simple Random Sampling (SRS) in e-Discovery. *Organizing Committee of the Fourth DESI Workshop on Setting Standards for Electronically Stored Information in Discovery*, 2011.
[8] Liang, J., Wang, Z., & Liu, X.. State estimation for coupled uncertain stochastic networks with missing measurements and time-varying delays: the discretetime case. *IEEE Transactions on Neural Network* 2000; 20, 5, 781–793.
[9] UCI machine learning repository. URL http://www.ics.uci.edu/~mlearn/MLRepository.html. 2011.
[10] Tahseen A. Jilani, Huda Yasin, Madiha Mohammad Yasin. PCA-ANN for Classification of Hepatitis Patients. *International Journal of Computer Applications* 2011; 14: 7.